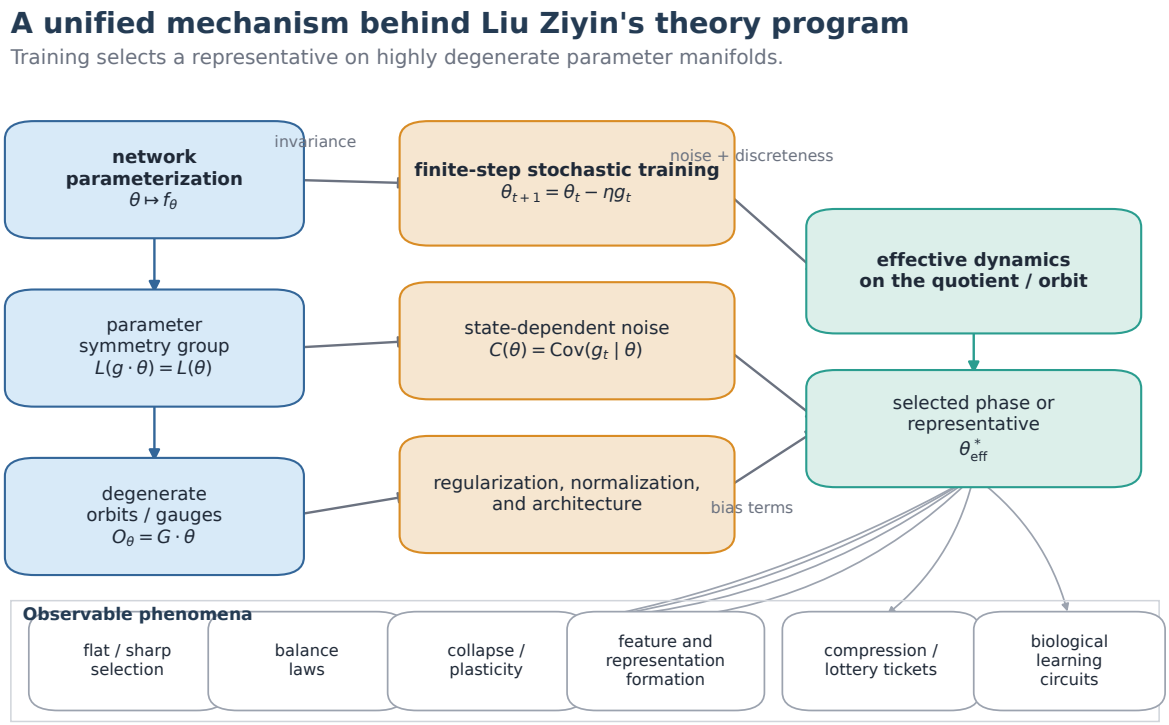
Liu Ziyin's Theory of Deep Learning

Symmetry, Stochastic Gradient Descent, Representations, and Neural Thermodynamics

Improved technical report with original explanatory figures



Prepared for:	the user
Prepared by:	OpenAI GPT-5.5 Pro
Version:	revised and reformatted, July 8, 2026
Figure status:	all figures are original explanatory schematics generated for this report

Abstract

Liu Ziyin's deep-learning theory program can be read as a single research agenda: modern networks are heavily overparameterized, symmetry-rich systems, and realistic training is a finite-step, stochastic, nonequilibrium process on those parameter spaces. The main results are not one theorem but a network of mechanisms. Finite learning rate changes stationary fluctuations; minibatch noise is anisotropic and state-dependent; SGD escape can be governed by logarithmic rather than ordinary loss barriers; parameter symmetries impose constraints, balance laws, Noether-like flows, and topology-preserving dynamics; learned representations tend toward canonical or polynomially aligned structures; collapse and feature learning can appear as phase transitions; and thermodynamic irreversibility reframes optimization as a physical process with entropic forces. This report reorganizes the papers by mechanism, explains the technical content with equations, and names major remaining open problems. Some 2025–2026 items are preprints or recent conference publications, so bibliographic details may change.

Contents

List of Figures

1 Executive synthesis

Thesis

The organizing claim is that deep learning is best modeled as *stochastic dynamics on a degenerate parameter manifold*. Ordinary empirical risk minimization specifies a level set of good functions; finite-step SGD, gradient noise, regularization, normalization, and parameter symmetries select a particular representative, phase, or quotient-space trajectory within that level set.

The public publication record on Liu Ziyin's homepage lists work spanning SGD noise, symmetry, representation formation, collapse, phase transitions, thermodynamics, compression, and biological learning [1, 2]. This report treats those papers as a coherent theory program rather than as isolated results. The most compact summary is:

$$\boxed{\text{symmetry + finite-step stochastic dynamics + regularization} \implies \text{implicit structure.}} \tag{1}$$

The structure may be a balanced parameterization, a low-rank or sparse solution, a collapsed representation, a canonical feature frame, a topology-preserving flow, a compressed subnet, or a biologically plausible local circuit.

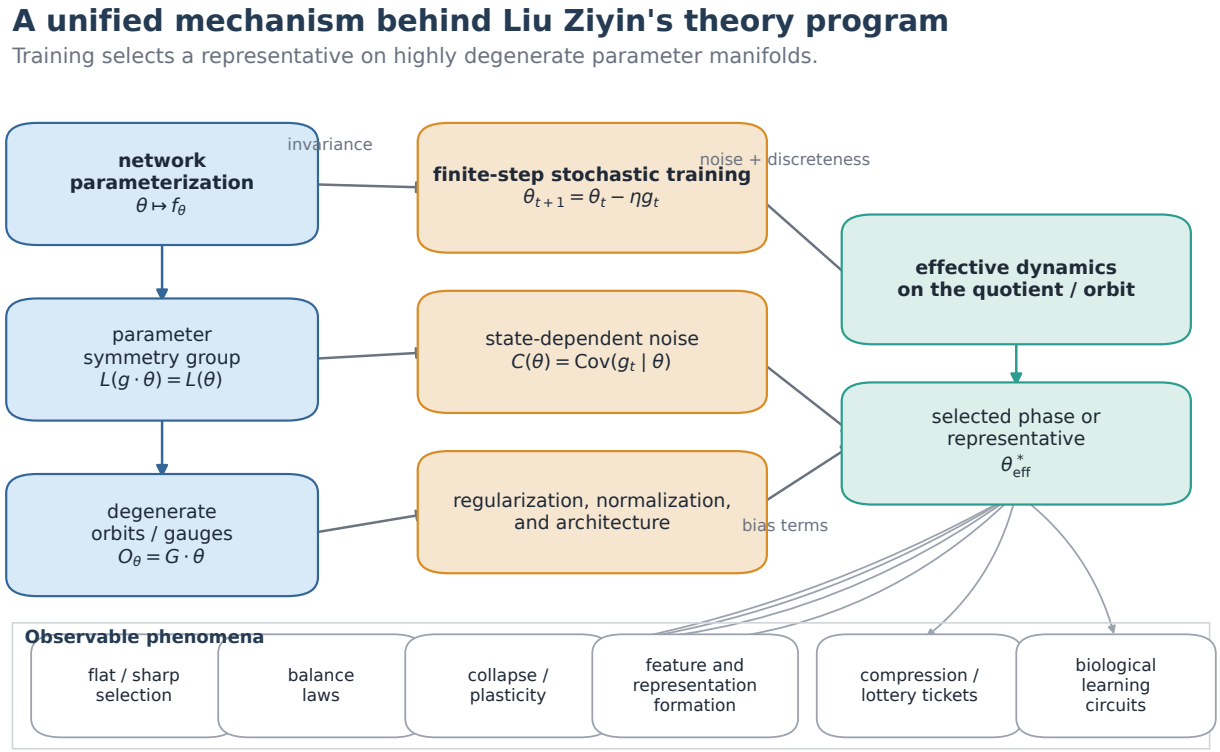


Figure 1: Original mechanism map. The report is organized around the pipeline shown here: neural parameterization gives symmetries and degeneracies; finite-step stochastic training creates effective dynamics on or near symmetry orbits; the selected representative explains the observed phenomena.

1.1 The main results in one page

Result family 1: finite-step SGD has its own stationary physics

For a quadratic basin, the exact discrete stationary covariance solves

$$\Sigma = (I - \eta H)\Sigma(I - \eta H)^\top + \eta^2 C, \tag{2}$$

not merely the continuous-time Lyapunov approximation $H\Sigma + \Sigma H \approx \eta C$. This difference is small only when $\eta\lambda_i(H) \ll 1$. Liu's SGD-noise papers show that many training phenomena live outside that small-step regime [6, 8].

Result family 2: flatness is not the primitive implicit bias

The 2026 exactly solvable model argues that SGD has no universal built-in preference for flat minima. It prefers small gradient fluctuations. Flatness emerges when gradient fluctuations and curvature are aligned, but anisotropic label noise can make the selected solution sharp [36].

Result family 3: parameter symmetries impose constraints

If $L(g \cdot \theta) = L(\theta)$, the loss alone cannot distinguish points on the orbit $G \cdot \theta$. Weight decay, gradient noise, discretization, or entropic corrections can nevertheless distinguish them. This yields sparsity, low rank, balance, homogeneous ensembling, or collapse, depending on the symmetry type [19, 20, 18].

Result family 4: representations form by alignment and symmetry breaking

The Canonical Representation Hypothesis says hidden representations, weights, and neuron gradients tend to align during training; when exact alignment fails, the Polynomial Alignment Hypothesis predicts power-law relations among the same objects [23]. This ties representation learning to dynamics rather than only to static architecture.

Result family 5: thermodynamic corrections explain effective forces

Neural thermodynamics interprets stochastic finite-step training as creating entropic forces that break some continuous reparameterization symmetries, preserve discrete or orthogonal ones, and select trajectories with lower entropy production [25, 37].

Result family 6: collapse and phase transitions are not accidents

In self-supervised learning, VAEs, deep linear networks, and exactly solvable feature-learning models, collapse can be a stable phase, and feature learning can occur through discontinuous or continuous transitions [14, 12, 13, 22].

1.2 What changed in this improved version

This revision changes three things relative to the earlier report. First, the organization is mechanism-first: each paper is placed inside a shared technical framework. Second, the figures are more scientific and professional: they use axes, equations, phase diagrams, trajectory plots, and explicit open-problem maps. Third, the open-problems section is expanded into named research problems with concrete success criteria.

2 Corpus, source status, and taxonomy

2.1 Scope

The core corpus is Liu Ziyin's public deep-learning-theory work from roughly 2019–2026, centered on the papers listed on his MIT publication page [2]. I include papers that are directly about theory, optimization, representation, collapse, thermodynamics, compression, or biological learning. I include algorithmic papers such as *LaProp*, *Snake*, *spred*, and *syre* because they are theory-motivated interventions. I exclude finance papers and most applied multimodal papers unless they directly bear on representation alignment.

Bibliographic caveat

Some works in 2025–2026 are preprints or recently accepted papers. Their title, venue, arXiv version, or proof details may change. The goal here is to explain the intellectual content and the stated main results, not to provide a final historical bibliography.

2.2 Chronology by technical lane

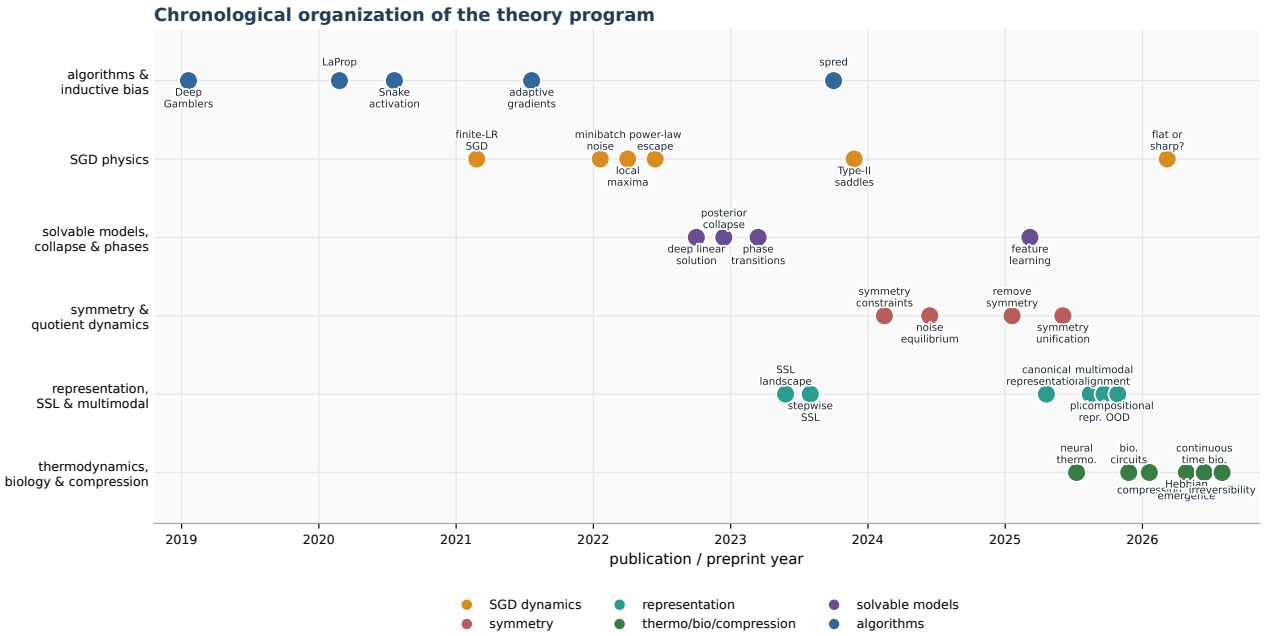


Figure 2: Original chronology organized by technical lane. The early work emphasizes optimizer behavior and SGD physics; the middle period develops exact models, collapse, and phase transitions; the recent period emphasizes symmetry, representation alignment, thermodynamics, compression, and biological learning.

2.3 Paper map

Table 1: Mechanism-first map of Liu Ziyin's major deep-learning-theory papers.

Year	Paper or paper group	Mathematical object	Main contribution
2019	*Deep Gamblers* [3]	Selective classification loss; portfolio analogy	Designs a loss that allows abstention under uncertainty. Peripheral to the later theory program, but it shows the style of importing structure from another mathematical domain.
2020	*LaProp* [4]	Adaptive optimizer update rule	Separates momentum and adaptivity to avoid an instability in Adam-style optimizers.
2020	*Neural Networks Fail to Learn Periodic Functions* [5]	Activation-function inductive bias	Shows standard activations extrapolate periodic functions poorly; proposes Snake, $x + \sin^2 x$.
2021	*Noise and Fluctuation of Finite Learning Rate SGD* [6]	Discrete stochastic recursion near minima	Shows finite learning rate changes the stationary distribution, momentum behavior, and escape compared with diffusion approximations.
2021	*Distributional Properties of Adaptive Gradients* [7]	Statistics of adaptive-gradient updates	Gives a distributional analysis of adaptive gradients and challenges a common explanation of Adam warmup.
2022	*Strength of Minibatch Noise in SGD* [8]	Per-example gradient covariance	Derives formulas for minibatch-noise strength near minima and links noise to stability, scaling, and implicit regularization.
2022	*SGD with a Constant Large Learning Rate Can Converge to Local Maxima* [9]	Discrete stochastic dynamics	Constructs examples showing large-learning-rate SGD can converge to local maxima, escape saddles slowly, or prefer sharp minima.
2022	*Logarithmic Landscape and Power-Law Escape Rate of SGD* [10]	Escape-rate theory	Replaces ordinary loss-barrier intuition with logarithmic landscape barriers and power-law escape rates.
2022–2023	*Exact Solutions of a Deep Linear Network* [11]	Deep linear network with regularization/stochasticity	Solves global minima and shows how deep parameterization creates bad minima at the origin and unusual degeneracies.
2022	*Posterior Collapse of a Linear Latent Variable Model* [12]	Linear VAE-like latent variable model	Identifies a precise likelihood-prior competition mechanism for posterior collapse.
2023	*Zeroth, First, and Second-Order Phase Transitions* [13]	Statistical-mechanics order parameters	Shows learning regimes can undergo first- or second-order phase transitions.
2023	*What Shapes the Loss Landscape of SSL?* [14]	Joint-embedding SSL loss	Characterizes complete and dimensional collapse, and the effects of normalization and bias.
2023	*On the Stepwise Nature of SSL* [15]	Eigenmodes of a contrastive kernel	Shows SSL can learn embedding dimensions sequentially, one mode at a time.
2023	*spred* [16]	Redundant reparameterization	Solves L_1-type sparse learning with differentiable SGD plus L_2-style decay.
2023	*Type-II Saddles* [17]	Random matrix products near saddles	Shows some saddles can be probabilistically stable for SGD because noise vanishes or aligns there.
2023–2025	*Law of Balance and Stationary Distribution of SGD* [18]	Rescaling-symmetric products	Studies balance, stationary measures, broken ergodicity, and fluctuation inversion in symmetry-rich models.
2024	*Symmetry Induces Structure and Constraint of Learning* [19]	Group actions and fixed subspaces	Proves that symmetry plus decay/noise can impose sparsity, low rank, or homogeneous ensembling.
2024	*Parameter Symmetry and Noise Equilibrium of SGD* [20]	Noether-like drift on symmetry orbits	Shows gradient noise can induce flow along symmetry directions until a noise equilibrium is reached.

Year	Paper or paper group	Mathematical object	Main contribution
2025	*Remove Symmetries to Control Model Expressivity* [21]	Symmetry-induced low-capacity states	Proposes syre, a model-agnostic method to break expressivity-limiting symmetries.
2025	*When Does Feature Learning Happen?* [22]	Analytically solvable finite-width model	Identifies alignment, disalignment, and rescaling as mechanisms absent from kernel limits.
2025	*Formation of Representations in Neural Networks* [23]	Alignment among R, W, and G	Proposes the Canonical Representation Hypothesis and Polynomial Alignment Hypothesis.
2025	*Parameter Symmetry Breaking and Restoration* [24]	Symmetry hierarchy	Frames hierarchical learning and representation formation as symmetry breaking and restoration.
2025–2026	*Neural Thermodynamics* [25] and *Thermodynamic Irreversibility* [37]	Entropic forces and entropy production	Recasts finite-step stochastic training as an irreversible nonequilibrium process with effective forces.
2025–2026	*Platonic Representation*, *multimodal alignment*, and compositionality papers [26, 27, 28]	Representation equivalence and factorization	Studies when representations align across architectures, modalities, and tasks, and when disentanglement is insufficient.
2025–2026	Topology, equivariance, compression [33, 34, 35]	Equivariant dynamics; permutation-invariant compression	Extends symmetry analysis to second-order geometry, topology preservation/breakdown, and lottery-ticket-style compression.
2024–2026	Biological learning papers [29, 30, 31, 32]	Heterosynaptic circuits and continuous-time learning	Argues that gradient-like learning may be implementable by biologically plausible motifs and temporal credit assignment.

3 Notation and technical background

Let $\theta \in \mathbb{R}^p$ denote network parameters and let $L(\theta)$ be the empirical loss. The basic stochastic update is

$$\theta_{t+1} = \theta_t - \eta g_t, \qquad g_t = \nabla L(\theta_t) + \xi_t, \qquad \mathbb{E}[\xi_t \mid \theta_t] = 0. \tag{3}$$

The conditional noise covariance is

$$C(\theta) = \mathbb{E}[\xi_t \xi_t^\top \mid \theta_t = \theta]. \tag{4}$$

In minibatch SGD, $C(\theta)$ is induced by sampling training examples. A finite-population approximation is

$$C_B(\theta) \approx \left(\frac{1}{B} - \frac{1}{N}\right) C_{\text{single}}(\theta), \tag{5}$$

where B is the minibatch size, N is the dataset size, and C_{single} is the covariance of per-example gradients. This formula is only a starting point: the important feature is that $C(\theta)$ is anisotropic and state-dependent.

A parameter symmetry is a transformation g such that

$$L(g \cdot \theta) = L(\theta), \qquad \text{or more strongly} \qquad f_{g\cdot\theta} = f_\theta. \tag{6}$$

Examples include hidden-unit permutations, rescaling of homogeneous layers, rotations of latent features, sign flips, and gauge symmetries in deep linear networks. Symmetry creates orbits

$$\mathcal{O}_\theta = \{g \cdot \theta : g \in G\} \tag{7}$$

on which the ordinary loss cannot choose a unique representative. Liu's theory asks which representative is selected by finite-step stochastic training.

4 Core result I: finite-step SGD is a physical stochastic process

4.1 Discrete covariance replaces diffusion covariance

Near a nondegenerate local minimum, approximate the loss by

$$L(\theta) = \frac{1}{2}\theta^\top H\theta. \tag{8}$$

With constant local noise covariance C, the exact stationary covariance of the discrete recursion is Eq. (2). The continuous-time diffusion approximation gives

$$H\Sigma + \Sigma H \approx \eta C. \tag{9}$$

The two equations agree to leading order only when ηH is small in every relevant direction. Modern training often deliberately approaches the edge of stability, so the discrete correction is not cosmetic.

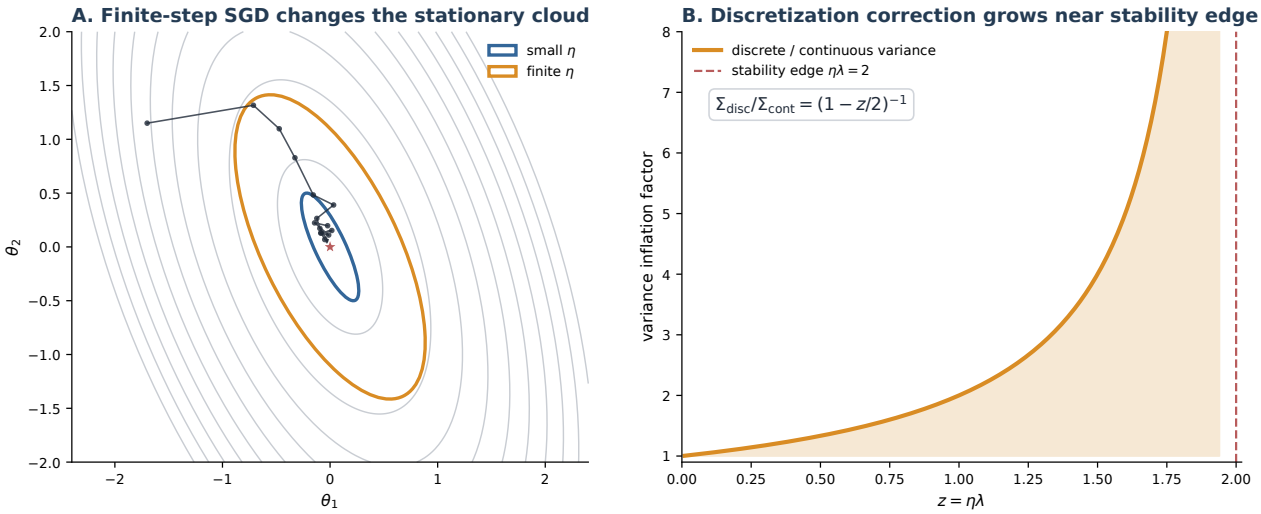


Figure 3: Original figure. Left: exact discrete SGD and the small-step approximation predict different stationary clouds in the same quadratic basin. Right: the variance-amplification factor grows sharply near the stability edge. This is the mathematical reason finite-learning-rate analyses are not reducible to gradient-flow analyses.

Noise and Fluctuation of Finite Learning Rate SGD studies this discrepancy and its extensions to momentum, minibatch noise, Adam-like methods, and escape [6]. The conceptual result is that the algorithm itself changes the stationary measure. The loss and Hessian do not fully specify the trained state.

4.2 Minibatch noise is structured

Strength of Minibatch Noise in SGD derives formulas for the magnitude of minibatch noise near minima and connects them to training stability, learning-rate/batch-size scaling, and implicit regularization [8]. The key point is that gradient noise is not isotropic "temperature." It is produced by per-example gradients, so it reflects data geometry and model representation. Consequently, any theory that treats SGD as isotropic Langevin noise will miss important effects.

The practical lesson is that stability depends on a joint object:

$$\text{stability and bias} \quad \approx \quad F\big(H(\theta), C(\theta), \eta, B, \text{optimizer state}\big). \tag{10}$$

This is a recurring theme in the later thermodynamic work.

4.3 Escape is governed by logarithmic barriers

Classical Kramers escape theory suggests a mean escape time like

$$\tau \sim \exp\left(\frac{\Delta L}{T}\right), \tag{11}$$

where ΔL is a loss barrier and T is an effective temperature. In *Logarithmic Landscape and Power-Law Escape Rate of SGD*, Liu and collaborators argue that SGD escape can instead be controlled by a logarithmic landscape, roughly by barriers in $\log L$ and by an effective Hessian dimension [10]. The output is a power-law escape rate rather than an ordinary Arrhenius exponential. This helps explain why SGD can move among basins in ways that are not captured by isotropic thermal diffusion.

4.4 Large-learning-rate SGD can violate gradient-flow intuition

SGD with a Constant Large Learning Rate Can Converge to Local Maxima constructs examples where finite-step stochastic updates converge to local maxima, escape saddles arbitrarily slowly, or prefer sharp minima [9]. These are not claims that all real training goes to maxima. They are impossibility results for a simple folk picture: SGD is not merely noisy descent.

Type-II Saddles and Probabilistic Stability of SGD makes this more precise near saddles. The local recursion can look like a random matrix product

$$\theta_{t+1} = A_{i_t}\theta_t, \tag{12}$$

where i_t indexes the sampled example or minibatch. Stability is governed by Lyapunov exponents of products of random matrices, not just by the eigenvalues of an averaged Hessian [17].

4.5 Flatness versus gradient fluctuations

The recent paper *Does SGD Seek Flatness or Sharpness?* gives a useful correction to a common slogan. In the solvable model, SGD selects minima with minimal gradient fluctuations; flatness is selected only when label noise is isotropic and fluctuation geometry agrees with curvature geometry [36]. When label noise is anisotropic, the same SGD criterion can select sharp solutions.

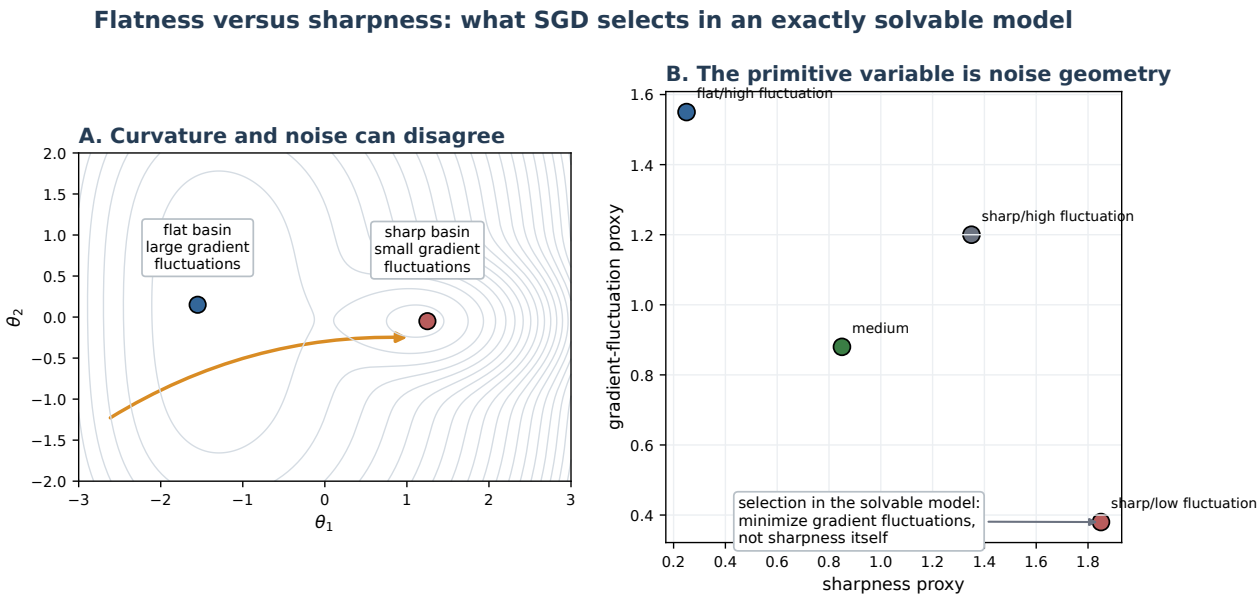


Figure 4: Original figure. The implicit bias is better stated in terms of gradient fluctuations than in terms of curvature alone. Flatness can be a proxy when fluctuation and curvature geometry align, but the proxy can fail.

Interpretation

The conclusion is not that flatness is irrelevant. It is that flatness is not the primitive variable in this line of theory. Curvature, noise covariance, stability, entropy production, and data distribution jointly determine the selected solution.

5 Core result II: parameter symmetry is the hidden structure

5.1 Symmetry-induced constraints

Symmetry Induces Structure and Constraint of Learning formalizes the idea that symmetries of the loss can impose structural constraints under weight decay or gradient noise [19]. Suppose M is a reflection symmetry:

$$L(M\theta) = L(\theta), \qquad M^2 = I. \tag{13}$$

The fixed subspace is

$$\mathcal{S} = \{\theta : M\theta = \theta\}. \tag{14}$$

A combination of symmetry, noise, and regularization can force learning toward such a subspace. The same logic changes form for other symmetry groups:

- rescaling symmetry can lead to balance or sparsity,
- rotation symmetry can lead to low-rank structure,
- permutation symmetry can lead to homogeneous ensembling or neuron collapse.

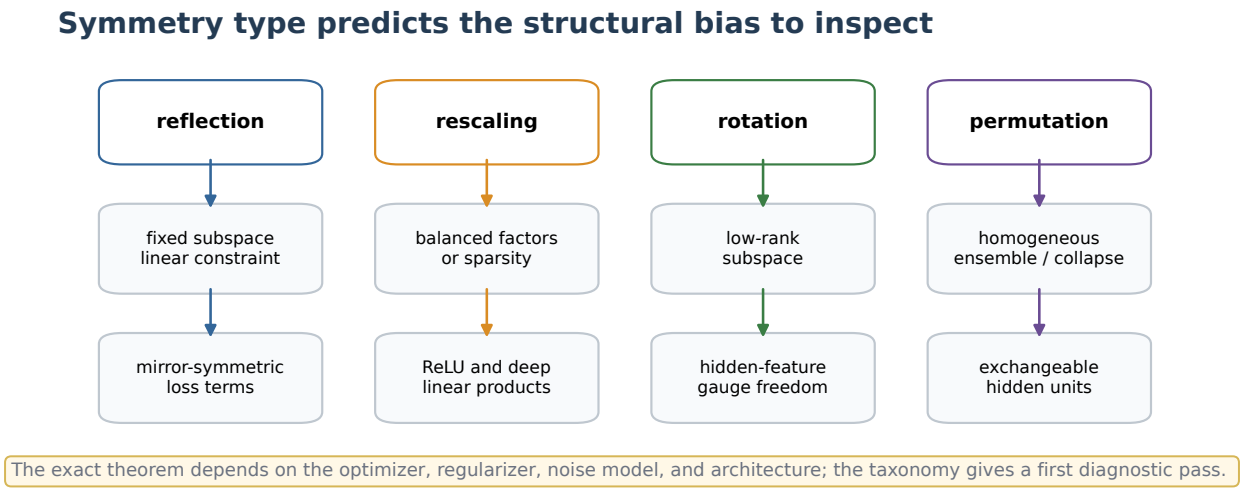


Figure 5: Original taxonomy. Different symmetry groups tend to imply different structural biases. The exact theorem depends on optimizer, regularization, noise model, and architecture, but the symmetry type tells you what to inspect.

5.2 Noise equilibrium and Noether flow

In a homogeneous two-layer model, a hidden unit often admits the rescaling

$$(a, w) \mapsto (c^{-1}a, cw), \qquad c > 0, \tag{15}$$

which preserves the represented function. The empirical loss is constant along the orbit. However, the gradient-noise covariance is generally not constant along that orbit. Therefore, stochastic training can drift along a direction in which the ordinary loss gradient is zero.

Parameter Symmetry and Noise Equilibrium of SGD formalizes this as a Noether-like flow induced by gradient noise [20]. The schematic statement is

$$\text{loss symmetry} + \text{nonuniform gradient noise} \implies \text{drift along symmetry orbit.} \tag{16}$$

The drift stops at a noise equilibrium, a special representative on the symmetry orbit.

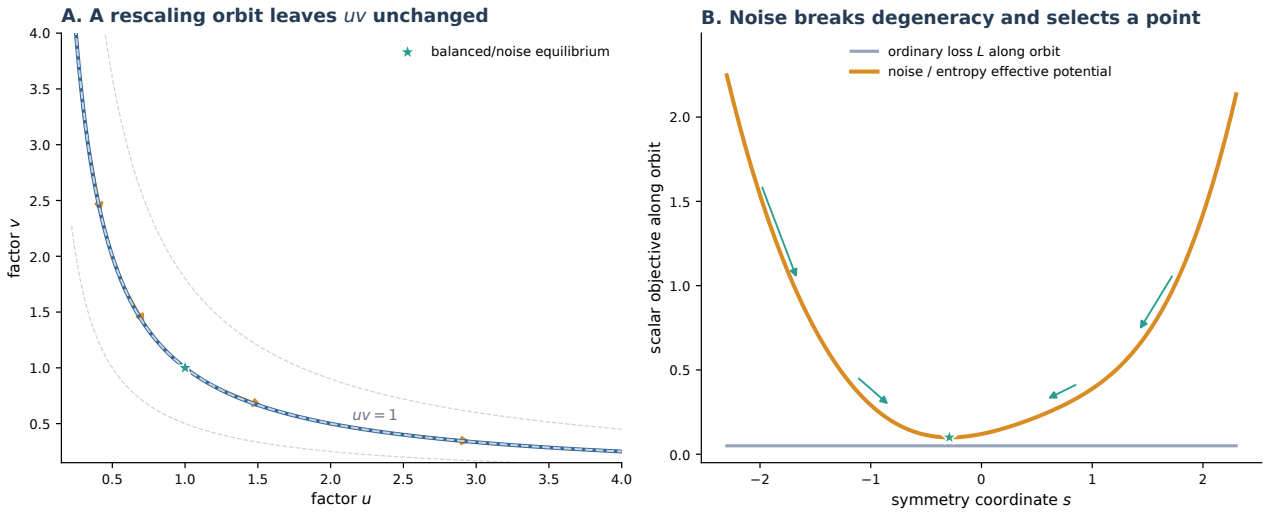


Figure 6: Original figure. Left: a simple product uv has a loss-invariant rescaling orbit $uv = 1$. Right: noise, decay, or entropic terms create an effective potential along the scale coordinate s, selecting a balanced representative.

5.3 Balance laws

For deep homogeneous models, balance usually means adjacent factors have matched norms or Gram matrices. In a simple product $y = uvx$, the function depends on uv but not on the individual scales. Balance means $|u| \approx |v|$. In matrix form, a common deep-linear balance condition is

$$W_\ell^\top W_\ell \approx W_{\ell+1} W_{\ell+1}^\top. \tag{17}$$

Law of Balance and Stationary Distribution of SGD studies stationary distributions in such symmetry-rich settings, including phase transitions, broken ergodicity, and fluctuation inversion [18]. The law of balance is not an arbitrary heuristic; it is the equilibrium selected by stochastic dynamics on a symmetry orbit.

5.4 Equivariance and topology

The 2025–2026 symmetry papers generalize this logic. *An Equivariance Toolbox for Learning Dynamics* derives first- and second-order constraints on gradients, Hessians, flat/sharp directions, and gradient-Hessian alignment from equivariance [34]. *Topological Invariance and Breakdown in Learning* studies how permutation-equivariant updates can preserve neuron-distribution topology below a critical learning rate and simplify it above [33]. In this view, learning rate is not only a speed parameter; it can control the topology of the representation trajectory.

6 Core result III: representations form through alignment

6.1 Canonical Representation Hypothesis

Let R denote hidden representations, W weights, and G neuron gradients. *Formation of Representations in Neural Networks* proposes the Canonical Representation Hypothesis (CRH): during learning, the subspaces associated with R, W, and G tend to align [23]. A stylized diagnostic is

$$A_{RW}(t) = \frac{\|P_R(t) P_W(t)\|_F}{\sqrt{\mathrm{rank}(P_R)}} \tag{18}$$

for projection matrices P_R and P_W. CRH predicts that such alignment statistics rise during training.

When exact alignment is impossible, the Polynomial Alignment Hypothesis (PAH) predicts reciprocal power-law relations among representations, weights, and gradients. This connects Liu's work to neural collapse and to the broader observation that trained features often become much more structured than random overparameterization would suggest.

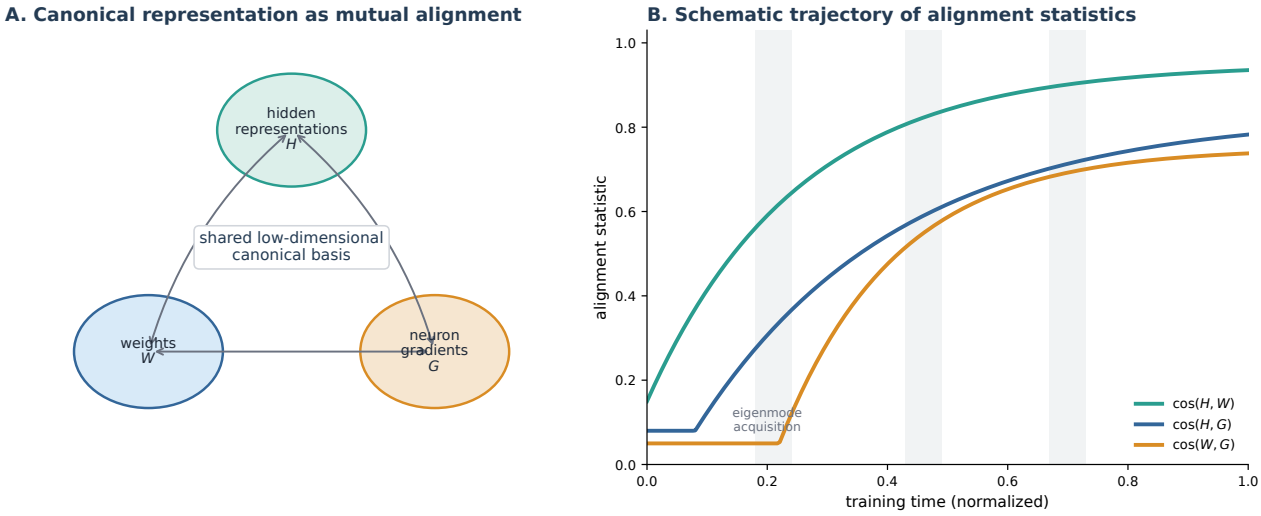


Figure 7: Original figure. CRH views representation formation as mutual alignment among hidden representations R, weights W, and neuron gradients G. The right panel shows schematic alignment statistics over normalized training time.

6.2 Platonic and multimodal alignment

The note *Proof of a Perfect Platonic Representation Hypothesis* proves, in a specific deep linear setting, that networks of different width, depth, and data can learn representations equivalent up to rotation [26]. This is an exact version of a broader empirical claim: sufficiently trained models often arrive at similar internal feature geometry.

Understanding the Emergence of Multimodal Representation Alignment studies when independently trained unimodal representations align [27]. The answer is data-dependent: redundant and shared information across modalities favors alignment, while unique modality-specific information limits it. This is important because it prevents overstatement. The theory is not that all good representations must be identical; it is that data structure plus training dynamics can create alignment when the task supports it.

6.3 Compositional generalization requires more than bottleneck disentanglement

Compositional Generalization Requires More Than Disentangled Representations shows that putting disentangled variables in a bottleneck is insufficient for robust out-of-distribution compositionality [28]. Later layers can re-entangle the factors, or the model can memorize through superposition. The broader message is consistent with CRH: representation quality is a property of the full learned dynamical system, not only of a single bottleneck layer.

7 Core result IV: solvable models expose collapse and phase transitions

7.1 Periodic functions and inductive bias

Neural Networks Fail to Learn Periodic Functions and How to Fix It shows that networks with common activations such as ReLU, tanh, and sigmoid can fail to extrapolate simple periodic functions [5]. The Snake activation

$$\phi(x) = x + \sin^2 x \tag{19}$$

injects periodic inductive bias while preserving a learnable nonperiodic component. This paper is not mainly about symmetry, but it illustrates a recurring methodological pattern: isolate a failure mode in a simple model, then introduce a targeted inductive bias.

7.2 Deep linear networks and the special role of zero

Exact Solutions of a Deep Linear Network solves global minima for a regularized deep linear network with stochastic neurons [11]. The key conceptual point is that deep parameterization makes the origin special. With more than one hidden layer, weight decay can create bad minima at zero. This is an example of a recurring theme: simple-looking deep networks have loss landscapes qualitatively different from shallow or convex models.

7.3 Posterior collapse

Posterior Collapse of a Linear Latent Variable Model identifies a mechanism for collapse in linear latent variable models and linear VAEs [12]. Posterior collapse arises from competition between likelihood fitting and prior-induced regularization of the latent mean. This exact model clarifies why collapse can be a stable optimum rather than a training accident.

7.4 Phase transitions and stepwise learning

Zeroth, First, and Second-Order Phase Transitions in Deep Neural Networks imports the language of statistical physics into deep learning: learned features can turn on continuously, discontinuously, or through abrupt jumps depending on model depth and regularization [13]. *On the Stepwise Nature of Self-Supervised Learning* shows that joint-embedding SSL methods can learn dimensions one eigenmode at a time [15]. *When Does Feature Learning Happen?* identifies alignment, disalignment, and rescaling as finite-width feature-learning mechanisms that disappear in the kernel regime [22].

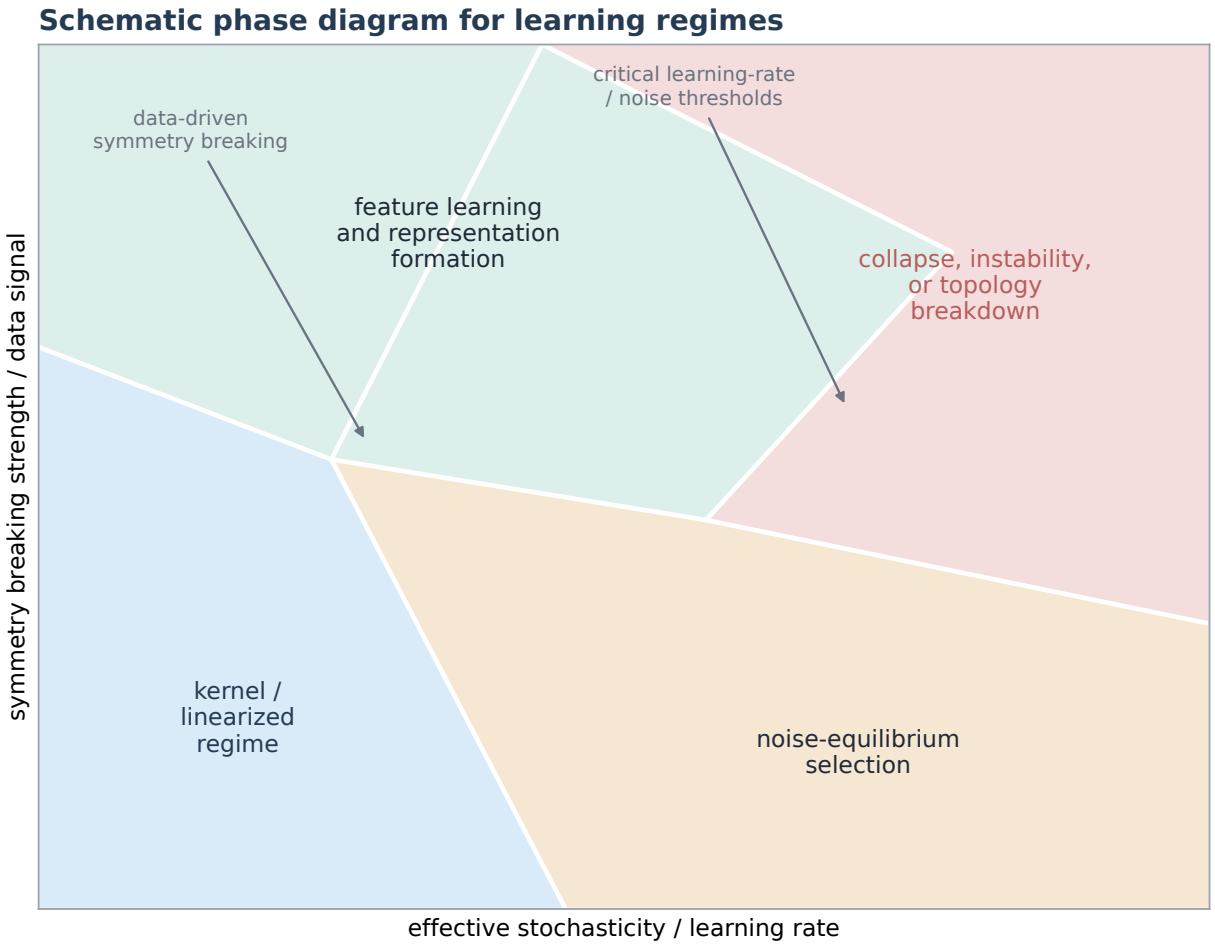


Figure 8: Original schematic phase diagram. The axes should not be read as universal coordinates; they summarize a repeated pattern in Liu's solvable models: changing learning rate, stochasticity, symmetry breaking, signal strength, width, or regularization can move a system among kernel-like, feature-learning, collapse, and noise-equilibrium regimes.

7.5 Self-supervised loss landscapes

What Shapes the Loss Landscape of Self-Supervised Learning? gives a tractable theory of joint-embedding SSL losses and characterizes complete collapse and dimensional collapse [14]. Normalization, bias, and the data spectrum can change whether collapse occurs. This is important because SSL collapse is often described as an engineering failure; Liu's work treats it as a predictable phase of the objective and dynamics.

8 Core result V: neural thermodynamics and irreversibility

8.1 Effective loss and entropic forces

Neural Thermodynamics argues that stochastic finite-step training creates emergent entropic forces [25]. A schematic way to write the idea is

$$L_{\text{eff}}(\theta) = L(\theta) + \eta S_1(\theta) + \eta^2 S_2(\theta) + \cdots , \tag{20}$$

where the correction terms depend on the optimizer, noise, discretization, and geometry. Equation (20) is not meant as a universal closed form; it is a template for why training can move along loss-invariant directions.

In this view, stochasticity is not merely error around gradient descent. It is a physical source of effective force. The theory claims that such forces systematically break non-isometric continuous parameter symmetries, preserve certain discrete or orthogonal symmetries, and lead to gradient-balance phenomena reminiscent of equipartition.

8.2 Thermodynamic irreversibility

Thermodynamic Irreversibility of Training Algorithms formalizes the irreversibility of training algorithms [37]. It relates four notions: numerical backward error ϕ_{DE}, time-renormalized correction ϕ_{TR}, microscopic time-reversal asymmetry ϕ_{TA}, and stochastic-thermodynamic entropy production ϕ_{ST}. The leading-order claim can be summarized as

$$\phi_{\mathrm{DE}} \simeq \phi_{\mathrm{TR}} \simeq \phi_{\mathrm{TA}} \simeq \phi_{\mathrm{ST}} \quad \text{to leading order in } \eta. \tag{21}$$

The irreversibility produces a time-reversal-symmetry-breaking emergent force and favors trajectories with lower entropy production.

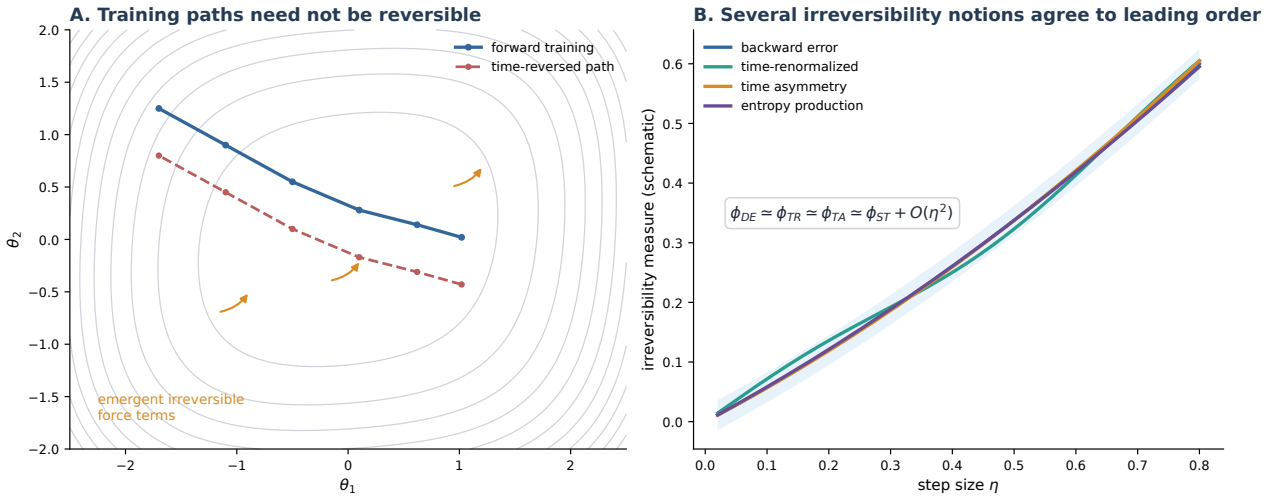


Figure 9: Original figure. Left: training paths need not be reversible. Right: different formal definitions of irreversibility agree to leading order in the step size in the thermodynamic irreversibility framework.

8.3 The thermodynamic synthesis

The thermodynamic papers make explicit what was implicit in the earlier SGD and symmetry papers: training is not a conservative optimization process. It has arrow-of-time structure. This matters because many deep-learning phenomena occur not at the minimizer set alone, but in the route by which the algorithm reaches that set. Trajectory-dependent quantities such as entropy production, noise covariance, and effective loss corrections become candidates for explaining representation formation and generalization.

9 Core result VI: theory-inspired interventions

9.1 Snake activation

Snake is an activation-function intervention motivated by a failure of ordinary activations on periodic extrapolation [5]. It is a local architectural change with a clear inductive-bias target.

9.2 LaProp and adaptive optimization

LaProp separates momentum and adaptivity in Adam-style methods [4]. Its role in the larger program is to show that optimizer details matter at the distributional level. The update rule is not an implementation detail; it changes the effective stochastic dynamics.

9.3 spred

spred: Sparsity by Redundancy turns L_1-regularized sparse learning into a differentiable redundant-parameter problem that can be solved with ordinary SGD and weight decay [16]. This is closely related to the symmetry program: redundant parameterization is not just overparameterization, but a way to convert nonsmooth structure into smooth dynamics.

9.4 syre

Remove Symmetries to Control Model Expressivity proposes syre, a method for escaping symmetry-induced low-capacity states [21]. The theory says a model can be expressive in principle while trapped inside a symmetric low-capacity submanifold in practice. Breaking the relevant symmetry can restore expressivity.

10 Core result VII: biological learning as gradient machinery

Liu's biological-learning papers ask whether gradient-like learning can be implemented by plausible circuits.

Self-Assembly of a Biologically Plausible Learning Circuit proposes a circuit motif capable of weight updates comparable to backpropagation in simulations and predicts a four-synapse motif between ascending and descending cortical streams [29]. *Heterosynaptic Circuits Are Universal Gradient Machines* argues that heterosynaptic plasticity can implement a broad class of gradient-based meta-learning [30]. *Emergence of Hebbian Dynamics in Regularized Non-Local Learners* shows that Hebbian or anti-Hebbian signatures can emerge from regularized nonlocal learning rules [31]. *On Biologically Plausible Learning in Continuous Time* studies continuous-time models that unify SGD, feedback alignment, direct feedback alignment, and Kolen-Pollack-like learning in limiting cases [32].

The connection to the rest of the theory is not merely metaphorical. In both artificial and biological settings, the question is how a local or finite-step dynamical system implements an effective gradient, and what extra forces or constraints appear because the implementation is not exact backpropagation.

11 Synthesis: a compact mathematical picture

The following hierarchy captures the program:

$$\text{Level 1: static objective} \quad L(\theta), \tag{22}$$

$$\text{Level 2: stochastic local geometry} \quad \big(H(\theta), C(\theta)\big), \tag{23}$$

$$\text{Level 3: symmetry structure} \quad G \curvearrowright \Theta, \quad L(g \cdot \theta) = L(\theta), \tag{24}$$

$$\text{Level 4: effective dynamics} \quad \theta_{t+1} = \theta_t - \eta \nabla L_{\text{eff}}(\theta_t) + \text{higher-order stochastic terms}, \tag{25}$$

$$\text{Level 5: selected phase} \quad \text{balanced, collapsed, aligned, compressed, or topology-changed state.} \tag{26}$$

This picture explains why the same vocabulary repeats across the papers:

- **Balance** is the selected representative on a rescaling orbit.

- **Collapse** is a stable symmetric or low-rank phase.
- **Flatness and sharpness** are consequences of fluctuation geometry, not universal primitives.
- **Canonical representation** is alignment among representation, weight, and gradient subspaces.
- **Thermodynamics** supplies the language for irreversible effective forces.
- **Compression and lottery tickets** exploit permutation symmetry and redundancy.

12 Major remaining open problems

The following open problems are phrased as research tasks. They are not criticisms of Liu's work; they are the natural next steps if the program is correct.

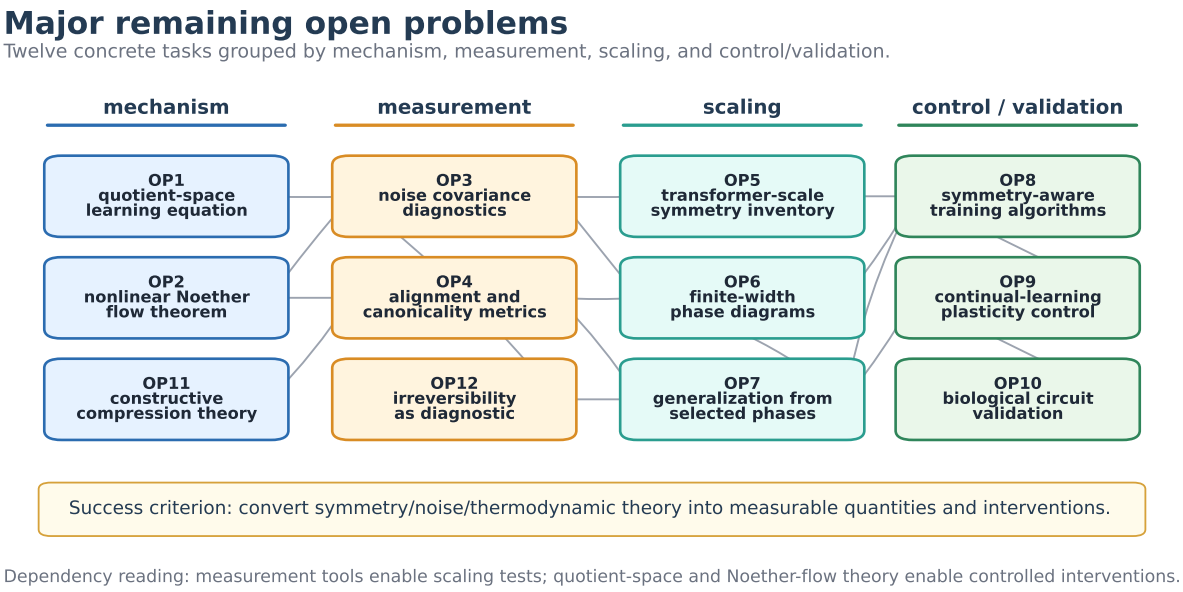


Figure 10: Original map of named open problems. The arrows indicate dependence: for example, transformer-scale symmetry inventory requires better diagnostics, and symmetry-aware training algorithms require quotient-space theory.

OP1. Quotient-space learning equation

Problem. Derive a practical, nonasymptotic learning equation on the quotient space Θ/G rather than on the full parameter space Θ.
Why it matters. If many observables are invariant under G, then the physically meaningful training state is not θ but an equivalence class $[\theta]$. A quotient-space equation would separate real function learning from gauge motion.
What would count as progress. A theorem or algorithm that predicts drift along symmetry orbits in a transformer block, convolutional network, or deep linear benchmark from measurable $C(\theta)$ and $H(\theta)$.

OP2. Nonlinear Noether-flow theorem

Problem. Generalize the Noether/noise-equilibrium theory from tractable settings to nonlinear, normalized, residual architectures with multiple interacting symmetries.
Why it is hard. In real networks, exact symmetries, approximate symmetries, batch/layer normalization, residual connections, weight decay, and adaptive optimizers interact. The symmetry group may not be cleanly identifiable.
What would count as progress. A theorem predicting which symmetry directions are broken, preserved, or balanced under SGD, AdamW, and normalized training.

OP3. Noise covariance diagnostics for large models

Problem. Estimate useful low-dimensional summaries of $C(\theta)$ during large-scale training.
Why it matters. Many results depend on gradient fluctuations, but exact per-example covariance is too expensive for frontier models.
What would count as progress. Scalable estimators for spectral norms, trace, low-rank components, anisotropy, and alignment between $C(\theta)$ and Hessian eigenspaces, with interventions that change predictions in the expected direction.

OP4. Alignment and canonicality metrics

Problem. Turn CRH and PAH into robust diagnostics that can be applied across architectures, datasets, and modalities.
Why it matters. Representation alignment is central to the theory, but empirical metrics can be sensitive to basis choices, normalization, layer selection, and data sampling.
What would count as progress. A benchmark suite where canonicality metrics predict transfer, robustness, SSL collapse, or multimodal alignment better than existing similarity measures alone.

OP5. Transformer-scale symmetry inventory

Problem. Catalog exact and approximate parameter symmetries in transformers, including attention heads, MLP channels, layer normalization, embeddings, residual streams, and attention logit shifts.
Why it matters. Liu's symmetry program is most powerful when the symmetry is explicit. Frontier architectures have many approximate symmetries but also many symmetry-breaking conventions.
What would count as progress. A symmetry atlas for transformer blocks plus empirical tests showing which symmetries control balance, collapse, head specialization, or feature superposition.

OP6. Finite-width phase diagrams

Problem. Build phase diagrams for realistic finite-width networks that interpolate among kernel learning, feature learning, collapse, edge-of-stability, and noise-equilibrium regimes.
Why it matters. Existing exact models show mechanisms, but practitioners need regime maps.
What would count as progress. Predictive boundaries in terms of width, learning rate, batch size, data spectrum, normalization, and optimizer, validated across multiple model classes.

OP7. Generalization from selected phases

Problem. Connect selected representatives or phases to test error, not only to training dynamics.
Why it is hard. A phase may be stable and interpretable without necessarily generalizing well. Generalization depends on data distribution, architecture, implicit bias, and evaluation task.

What would count as progress. A theorem or empirical law linking noise equilibrium, entropy production, canonicality, compression, or phase-transition order to out-of-sample performance.

OP8. Symmetry-aware training algorithms

Problem. Design optimizers and architectural interventions that intentionally control symmetry breaking, balance, collapse, and representation alignment.
Why it matters. syre and spred are early examples. A general toolbox would make the theory operational.
What would count as progress. Algorithms that diagnose a harmful symmetry-induced low-capacity state and automatically apply a targeted symmetry-breaking or symmetry-restoring intervention.

OP9. Continual-learning plasticity control

Problem. Extend the symmetry/noise theory to nonstationary, continual, reinforcement, and curriculum learning settings.
Why it matters. Loss of plasticity is closely related to symmetry-induced collapse and low-dimensional trapping, but most exact theory assumes stationary supervised datasets.
What would count as progress. A measurable criterion predicting when a trained model has entered a low-plasticity phase, plus interventions that restore plasticity without catastrophic forgetting.

OP10. Biological circuit validation

Problem. Test whether predicted heterosynaptic motifs and temporal overlap requirements appear in neural systems.
Why it matters. The biological-learning papers make stronger claims than analogies: they propose circuit mechanisms.
What would count as progress. Experiments that confirm or refute the predicted four-synapse motifs, heterosynaptic gradient computation, and plasticity-window constraints.

OP11. Compression theorem to constructive algorithm

Problem. Convert universal compression and dynamical lottery-ticket theory into practical compression algorithms for real networks and datasets.
Why it matters. The compression theory is powerful but relies on assumptions such as smoothness and permutation invariance. The practical challenge is to find the compressed object efficiently.
What would count as progress. A method that compresses networks or datasets by more than standard pruning/distillation baselines while preserving training dynamics, not only final outputs.

OP12. Irreversibility as a training diagnostic

Problem. Measure entropy production or equivalent irreversibility in real training and use it to predict learning quality.
Why it matters. Thermodynamic irreversibility is a unifying theoretical idea, but it becomes practically useful only if it yields measurable signals.
What would count as progress. A low-cost estimator of irreversibility that predicts phase changes, representation formation, optimizer instability, or generalization across tasks.

13 Recommended reading paths

13.1 Fast conceptual route

Read these in order:

1. *Symmetry Induces Structure and Constraint of Learning* [19].
2. *Parameter Symmetry and Noise Equilibrium of SGD* [20].
3. *Formation of Representations in Neural Networks* [23].
4. *Neural Thermodynamics* [25].
5. *Thermodynamic Irreversibility of Training Algorithms* [37].

This route gives the shortest path from symmetry to thermodynamics.

13.2 SGD physics route

Read:

1. *Noise and Fluctuation of Finite Learning Rate SGD* [6].
2. *Strength of Minibatch Noise in SGD* [8].
3. *Logarithmic Landscape and Power-Law Escape Rate of SGD* [10].
4. *Type-II Saddles and Probabilistic Stability of SGD* [17].
5. *Does SGD Seek Flatness or Sharpness?* [36].

This route explains why SGD must be analyzed as a finite-step stochastic process.

13.3 Representation route

Read:

1. *What Shapes the Loss Landscape of SSL?* [14].
2. *On the Stepwise Nature of SSL* [15].
3. *When Does Feature Learning Happen?* [22].
4. *Formation of Representations in Neural Networks* [23].
5. *Proof of a Perfect Platonic Representation Hypothesis* [26].

This route focuses on hidden-layer geometry and feature formation.

13.4 Solvable-model route

Read:

1. *Neural Networks Fail to Learn Periodic Functions* [5].
2. *Exact Solutions of a Deep Linear Network* [11].
3. *Posterior Collapse of a Linear Latent Variable Model* [12].
4. *Zeroth, First, and Second-Order Phase Transitions* [13].
5. *When Does Feature Learning Happen?* [22].

This route is best if you want exact toy models before reading the broad synthesis papers.

14 Critical assessment

14.1 What is strongest

The strongest part of Liu's program is mechanism-level analysis. The work repeatedly takes a vague empirical claim and replaces it with a tractable mathematical object:

- "SGD likes flat minima" becomes a statement about gradient-fluctuation geometry.
- "Overparameterization helps" becomes a statement about symmetry orbits and redundant parameterizations.
- "Features emerge" becomes an alignment problem among R, W, and G.
- "Collapse happens" becomes a stable phase of an objective and training dynamics.
- "Training is optimization" becomes a nonequilibrium thermodynamic process.

This is scientifically valuable because it produces falsifiable diagnostics and target quantities.

14.2 What is still speculative

The broadest unification claims are plausible but not yet complete. Exact calculations are strongest in simplified models. General transformers, AdamW, layer normalization, data augmentation, curriculum effects, and large-scale pretraining still require more direct quantitative theory. The open problems above are where the program must become more predictive.

14.3 Bottom line

Liu Ziyin's contribution is best described as a theory style: identify the parameter symmetry; measure the stochastic finite-step correction; derive the selected representative or phase; then test whether that representative explains collapse, balance, representation alignment, compression, or biological implementation. The value of the program is not that it solves all of deep learning theory, but that it gives a common mechanism for phenomena that otherwise look unrelated.

Source and figure note

This report was prepared from public web pages and public arXiv/OpenReview-style records visible as of July 8, 2026. All diagrams are original explanatory figures created for this report and are not copied from Liu's papers. The figures are schematic unless explicitly derived from the displayed toy equations; they are meant to clarify mechanisms, not to reproduce experimental results.

References

[1] Liu Ziyin. Personal website. https://www.mit.edu/~ziyinl/. Accessed July 8, 2026.

[2] Liu Ziyin. Publications page. https://www.mit.edu/~ziyinl/publication.html. Accessed July 8, 2026.

[3] Ziyin Liu, Zhikang Wang, Paul P. Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. *Deep Gamblers: Learning to Abstain with Portfolio Theory*. NeurIPS, 2019. https://arxiv.org/abs/1907.00208.

[4] Ziyin Liu, Zhikang Wang, and Masahito Ueda. *LaProp: Separating Momentum and Adaptivity in Adam*. 2020. https://arxiv.org/abs/2002.04839.

[5] Liu Ziyin, Tilman Hartwig, and Masahito Ueda. *Neural Networks Fail to Learn Periodic Functions and How to Fix It*. NeurIPS, 2020. https://arxiv.org/abs/2006.08195.

[6] Kangqiao Liu, Liu Ziyin, and Masahito Ueda. *Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent*. ICML, 2021. https://arxiv.org/abs/2012.03636.

[7] Zhang Zhiyi and Liu Ziyin. *On the Distributional Properties of Adaptive Gradients*. UAI, 2021. https://arxiv.org/abs/2105.07222.

[8] Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. *Strength of Minibatch Noise in SGD*. ICLR, 2022. https://arxiv.org/abs/2102.05375.

[9] Liu Ziyin, Botao Li, James B. Simon, and Masahito Ueda. *SGD with a Constant Large Learning Rate Can Converge to Local Maxima*. ICLR, 2022. https://arxiv.org/abs/2107.11774.

[10] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. *Logarithmic Landscape and Power-Law Escape Rate of SGD*. ICML, 2022. https://arxiv.org/abs/2105.09557.

[11] Liu Ziyin, Botao Li, and Xiangming Meng. *Exact Solutions of a Deep Linear Network*. NeurIPS, 2022; Journal of Statistical Mechanics: Theory and Experiment, 2023. https://arxiv.org/abs/2202.04777.

[12] Zihao Wang and Liu Ziyin. *Posterior Collapse of a Linear Latent Variable Model*. NeurIPS, 2022. https://arxiv.org/abs/2205.04009.

[13] Liu Ziyin and Masahito Ueda. *Zeroth, First, and Second-Order Phase Transitions in Deep Neural Networks*. Physical Review Research, 2023. https://arxiv.org/abs/2205.12510.

[14] Liu Ziyin, Ekdeep Singh Lubana, Masahito Ueda, and Hidenori Tanaka. *What Shapes the Loss Landscape of Self-Supervised Learning?* ICLR, 2023. https://arxiv.org/abs/2210.00638.

[15] James B. Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J. Fetterman, and Joshua Albrecht. *On the Stepwise Nature of Self-Supervised Learning*. ICML, 2023. https://arxiv.org/abs/2303.15438.

[16] Liu Ziyin and Zihao Wang. *Sparsity by Redundancy: Solving L_1 with SGD*. ICML, 2023. https://arxiv.org/abs/2210.01212.

[17] Liu Ziyin, Botao Li, and collaborators. *Type-II Saddles and Probabilistic Stability of Stochastic Gradient Descent*. 2023. https://arxiv.org/abs/2303.13093.

[18] Liu Ziyin, Hongchao Li, and Masahito Ueda. *Law of Balance and Stationary Distribution of Stochastic Gradient Descent*. Physical Review E. https://arxiv.org/abs/2308.06671.

[19] Liu Ziyin. *Symmetry Induces Structure and Constraint of Learning*. ICML, 2024. https://arxiv.org/abs/2309.16932.

[20] Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. *Parameter Symmetry and Noise Equilibrium of Stochastic Gradient Descent*. NeurIPS, 2024. https://arxiv.org/abs/2402.07193.

[21] Liu Ziyin, Yizhou Xu, and Isaac Chuang. *Remove Symmetries to Control Model Expressivity*. ICLR, 2025. https://arxiv.org/abs/2408.15495.

[22] Yizhou Xu and Liu Ziyin. *When Does Feature Learning Happen? Perspective from an Analytically Solvable Model*. ICLR, 2025. https://arxiv.org/abs/2401.07085.

[23] Liu Ziyin, Isaac Chuang, Tomer Galanti, and Tomaso Poggio. *Formation of Representations in Neural Networks*. ICLR, 2025. https://arxiv.org/abs/2410.03006.

[24] Liu Ziyin, Yizhou Xu, Tomaso Poggio, and Isaac Chuang. *Parameter Symmetry Breaking and Restoration Determines the Hierarchical Learning in AI Systems*. 2025. https://arxiv.org/abs/2502.05300.

[25] Liu Ziyin, Yizhou Xu, and Isaac Chuang. *Neural Thermodynamics: Entropic Forces in Deep and Universal Representation Learning*. NeurIPS, 2025. https://arxiv.org/abs/2505.12387.

[26] Liu Ziyin and Isaac Chuang. *Proof of a Perfect Platonic Representation Hypothesis*. 2025. https://arxiv.org/abs/2507.01098.

[27] Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. *Understanding the Emergence of Multimodal Representation Alignment*. ICML, 2025. https://arxiv.org/abs/2502.16282.

[28] Qiyao Liang, Daoyuan Qian, Liu Ziyin, and Ila Fiete. *Compositional Generalization Requires More Than Disentangled Representations*. ICML, 2025. https://arxiv.org/abs/2501.18797.

[29] Qianli Liao, Liu Ziyin, Yulu Gan, Brian Cheung, Mark Harnett, and Tomaso Poggio. *Self-Assembly of a Biologically Plausible Learning Circuit*. 2024. https://arxiv.org/abs/2412.20018.

[30] Liu Ziyin, Isaac Chuang, and Tomaso Poggio. *Heterosynaptic Circuits Are Universal Gradient Machines*. 2025. https://arxiv.org/abs/2505.02248.

[31] David Koplow, Tomaso Poggio, and Liu Ziyin. *Emergence of Hebbian Dynamics in Regularized Non-Local Learners*. ICML, 2026. https://arxiv.org/abs/2505.18069.

[32] Marc Gong Bacvanski, Liu Ziyin, and Tomaso Poggio. *On Biologically Plausible Learning in Continuous Time*. 2026. https://arxiv.org/abs/2510.18808.

[33] Yongyi Yang, Tomaso Poggio, Isaac Chuang, and Liu Ziyin. *Topological Invariance and Breakdown in Learning*. 2025. https://arxiv.org/abs/2510.02670.

[34] Yongyi Yang and Liu Ziyin. *An Equivariance Toolbox for Learning Dynamics*. 2025. https://arxiv.org/abs/2512.21447.

[35] Hong-Yi Wang, Di Luo, Tomaso Poggio, Isaac L. Chuang, and Liu Ziyin. *A Universal Compression Theory: Lottery Ticket Hypothesis and Superpolynomial Scaling Laws*. ICLR, 2026. https://arxiv.org/abs/2510.00504.

[36] Yizhou Xu, Pierfrancesco Beneventano, Isaac Chuang, and Liu Ziyin. *Does SGD Seek Flatness or Sharpness? An Exactly Solvable Model*. 2026. https://arxiv.org/abs/2602.05065.

[37] Liu Ziyin, Yuanjie Ren, Adam Levine, and Isaac Chuang. *Thermodynamic Irreversibility of Training Algorithms*. 2026. https://arxiv.org/abs/2605.21933.